ARL-TR-8124 • Sep 2017

US Army Research Laboratory

The Effect of Training Data Set Composition on the Performance of a Neural Image Caption Generator

by Abigail Wilson and Adrienne Raglin

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-8124 ● SEP 2017

US Army Research Laboratory

The Effect of Training Data Set Composition on the Performance of a Neural Image Caption Generator

by Abigail Wilson
Montgomery Blair High School

Adrienne Raglin
Computational and Information Sciences Directorate, ARL

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YYYY)* September 2017	**2. REPORT TYPE** Technical Report	**3. DATES COVERED (From - To)**

4. TITLE AND SUBTITLE The Effect of Training Data Set Composition on the Performance of a Neural Image Caption Generator	**5a. CONTRACT NUMBER**
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Abigail Wilson and Adrienne Raglin	**5d. PROJECT NUMBER**
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) US Army Research Laboratory ATTN: RDRL-CII-B 2800 Powder Mill Road Adelphi, MD 20783-1138	**8. PERFORMING ORGANIZATION REPORT NUMBER** ARL-TR-8124
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	**10. SPONSOR/MONITOR'S ACRONYM(S)**
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This research seeks to determine how many images of a particular object in a training data set are necessary to achieve caption quality saturation in neural image caption generators. Understanding the relationship between caption quality and the size and composition of training data sets could improve efficiency in model training and lead to the development of optimized data sets for different tasks. We hypothesize that increasing the exposure of a neural network to an object will improve its performance, up to a point, after which the caption quality will saturate; and that this may vary based on the object's visual homogeneity. We trained several image captioning models, using an existing code Neuraltalk2, on subsets of the Microsoft Common Objects in Context data set, which contained a precise number of some common object categories (e.g., “cat” and “pizza”). The performance with different levels of exposure to the selected objects was compared using the Metric for Evaluation of Translation with Explicit Ordering (METEOR) and Consensus-Based Image Description Evaluation (CIDEr) automated scoring metrics. The data indicate that increasing the quantity of images of a particular object in the training data set improved the performance up to 1,500 images, but not beyond that.

15. SUBJECT TERMS

neural image captions, training, storytelling, Microsoft Common Objects in Context Data set, MS COCO

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Adrienne Raglin
a. REPORT Unclassified	**b. ABSTRACT** Unclassified	**c. THIS PAGE** Unclassified	UU	20	**19b. TELEPHONE NUMBER (Include area code)** 301-394-0210

Contents

List of Figures

List of Tables

Acknowledgments

Abigail Wilson thanks Dr Adrienne Raglin for her mentorship and support, and Mr Michael Lee for his guidance throughout the project. She also expresses appreciation for Mr Michael Landes, Dr Andre Harrison, and Mr Chris Schlesiger for their technical assistance.

INTENTIONALLY LEFT BLANK.

1. Introduction

For computers, the task of generating captions for images is complex, as it requires both computer vision and natural language processing techniques. Computer-based models must understand what elements are contained in the image and how to effectively use language to describe them.[1] Traditionally, only humans have been able to caption images well because of the human ability to easily understand and provide interpretations of visual data.[2] However, machine learning is now being applied to this activity. Recent research has led to the creation of effective image captioning models using neural networks. In general, neural networks are systems used in machine learning that emulate the processes of biological neural networks to train models to complete certain tasks.

Most neural image caption generators are trained on data sets composed of thousands of images that have been annotated with human-generated captions. Several such data sets exist: Microsoft Common Objects in Context (MS COCO), Pascal VOC, Flickr 8k, Flickr 30k, and a few others. These data sets in particular contain between 8,000 and 80,000 images annotated with 5 captions each.[2,3] Additional resources are being devoted to generating larger data sets with different types of images and different levels of annotations. The literature describes the use of various forms of annotations and iconic versus noniconic images in training data sets, but we have not found research into how image type and quantity affect the performance of trained image-captioning models.

The performance of neural image caption generators can be measured using automated scoring systems. Many such systems exist, including Bilingual Evaluation Understudy (BLEU), Consensus-Based Image Description Evaluation (CIDEr), and the Metric for Evaluation of Translation with Explicit Ordering (METEOR). These metrics all compare reference captions with generated candidate captions. BLEU focuses on modified n-gram precision, but does not address recall.[4] This is seen as a drawback to BLEU, as recall has been shown to be essential to automated scoring, which correlates highly with human precision.[5] CIDEr uses a system of consensus among the captions and was created specifically for image captioning.[6] METEOR incorporates synonyms of target words using data from WordNet.[5] These metrics have been shown to agree with human judgment.[1,6]

The current standard in neural image captioning is the use of end-to-end neural networks to generate captions from an image input. This method was used both by Vinyals et al. and Karpathy and Li.[1,2] In these studies, the image input was fed to a Convolutional Neural Network (CNN) and the output was fed into a form of a Recurrent Neural Network (RNN) for caption generation.[1,2] This technique uses a unified process, which converts smoothly from images to captions. Both of these models performed well when evaluated using the BLEU-4 metric.

The CNN to RNN method, which both of those papers described, is the basis for the NeuralTalk code created by Dr Andrej Karpathy.[7] NeuralTalk is an open-source image-captioning program that allows the user to train new models on training data sets that contain annotated images.[7] We used the NeuralTalk2 code—an updated version of NeuralTalk—in this project.[8] NeuralTalk2 was written in Lua and runs in Torch. It improves upon NeuralTalk by making use of a graphical processing unit (GPU) to greatly decrease runtime.[8]

Neural image captioning is a relatively new field, and as such, there is limited research on the effect of different training data set compositions. The current trend is to create new data sets of ever-increasing size, but larger data sets lead to longer training times, extending what is already a resource-intensive process. Understanding the effect of the training data set composition is essential to moving toward more efficient, focused training. Currently, training a model can take days or even weeks. To optimize performance and quality in caption generation, it is important to know how the quantity of images in a training data set affects the resulting caption quality.

In this study, we hypothesized that as the quantity of images of a particular object increases, the caption quality will also increase. The hypothesis was modified with the caveat that there exists a point at which caption quality becomes saturated. However, this point may vary for objects with different levels of visual homogeneity as well as specificity in object types. For example, more images of dogs would be needed than soccer balls to achieve the same caption quality. The visual characteristics of different types of dogs are broad, while different soccer balls appear similar even from different viewpoints. The goal of the experiment for this work was to investigate the relationship between the number of exposures to a model of an object type and the model's performance on images including that object. In this report, the initial experimentation and results are presented.

2. Methods

2.1 Model

For this investigation, NeuralTalk2, a caption-generating program by Dr Andrej Karpathy of Stanford University, was used.[8] The code is publicly available on GitHub and runs in Torch. NeuralTalk2 is an improved version of NeuralTalk, which was written based on the models described earlier.[1,2] The program was designed for training new image-captioning models on image data sets.

NeuralTalk2 was chosen for its ease of use to focus the investigation on training and the potential impact on caption quality. Thus, to maximize the resources devoted to the training of the models, it was most efficient to work with an existing caption generator.

The setup included the NeuralTalk2 caption-generating program on an Ubuntu 14.04 operating system. Without access to a Compute Unified Device Architecture (CUDA) compatible NVidia GPU, the program was run in CPU-only mode. For this initial investigation, default parameters were used in training with the exception of the batch size, which was changed to 15 so that the program would use all of the validation images exactly once.

2.2 Data Set

The 2014 training and validation images from the MS COCO image data set were used as training material in this experiment. The MS COCO data set consists of over 80,000 training images and 40,000 validation and testing images.[3] Each image is annotated with 5 human-generated captions and bounding boxes. The bounding boxes outline objects in the image falling into one of the 80 object categories contained by the MS COCO data set. The training images used in this experiment were randomly pulled from the training image data set and the validation and testing images both came from the validation set.

The MS COCO data set's size and preexisting categories make it optimal for creating subsets with distinct composition. The large size of the MS COCO data set allowed the *parsing* of the data set into smaller training sets, each with particular characteristics, while retaining a sizable training set. The categories were used to focus on a few specific objects for this study.

The MS COCO data set also has an existing application programing interface (API).[9] The existing annotations and parsing actions provided in the software were used, greatly simplifying the process.

2.3 Experiment

To determine the relationship between the quantity of images of a particular object in a training data set and the performance of a caption-generating model, 6 object categories from the MS COCO data set were chosen. The number of instances of each of the 80 categories were calculated, focusing on the categories with over 2,000 instances. Limited descriptors were used. For example, "person" was excluded due to the use of other words to describe a person (e.g., "man", "girl", "doctor"). This simplified the number of images of people in the training data set controlling some of the variability within this experimental design. Categories “sink”, “cat”, “car”, “pizza”, “skateboard”, and “dog” were sample descriptors used for additional simplification.

To harness the MS COCO API, a program was written in Python. The program created subsets of the MS COCO data set with set numbers of images in specified categories. These subsets were outlined in JavaScript Object Notation (JSON) files describing the images and captions, and formatted for compatibility with the NeuralTalk2 data set preparation programs.

Using this program, training data sets were created. The training data sets were composed of a baseline set of 14,000 randomly chosen images that did not overlap with the previously chosen categories and a set of 6,000 images from the descriptor categories. The set of 6,000 images was structured to vary the quantity of images by category: one set contained 1,000 training images of each category, while another contained 500 images each of cats, cars, and dogs; and 1,500 each of sinks, pizzas, and skateboards. All overlapping images that fell into 2 or more categories were removed to maintain constant image totals and data set size. Each data set also contained 1,500 validation images that were distinct from the categories and 250 validation images from each category, for a total of 3,000 validation images. These validation images were constant across all models. In this fashion, 5 data sets were created each containing the same testing and validation images and the same total number of training images (Table 1). A control data set was also created, which contained the 14,000 base training images and the 1,500 base validation images. All models were tested on a constant set of 1,000 images, which were not from the 6 categories, and sets of 500 images from each category.

Table 1 The composition of all of the training data sets for the 6 models. Values represent the number of images of a particular category in the data set.

	Training Images								Validation Images			Testing images		
	Baseline	Cat	Car	Dog	Pizza	Sink	Skateboard	Total	Baseline	Categories	Total	Baseline	Categories	Total
Control	14,000	---	---	---	---	---	---	14,000	1,500	---	1,500	1,000	---	1,000
1	14,000	0	0	0	2,000	2,000	2,000	20,000	1,500	250 each	3,000	1,000	500 each	4,000
2	14,000	500	500	500	1,500	1,500	1,500	20,000	1,500	250 each	3,000	1,000	500 each	4,000
3	14,000	1,000	1,000	1,000	1,000	1,000	1,000	20,000	1,500	250 each	3,000	1,000	500 each	4,000
4	14,000	1,500	1,500	1,500	500	500	500	20,000	1,500	250 each	3,000	1,000	500 each	4,000
5	14,000	2,000	2,000	2,000	0	0	0	20,000	1,500	250 each	3,000	1,000	500 each	4,000

The data set specifications were written to JSON files. These files were processed using the prepro.py program from NeuralTalk2. Models were then trained on each of the training data sets using a batch size of 15 and all of the validation images. Each model trained for 20,000 iterations with no fine-tuning.

2.4 Metrics

The caption quality was measured with the automated scoring systems CIDEr[1] and METEOR[2]. These scores were calculated using the coco-caption code provided by Microsoft for evaluating captions.[9] In recent papers, CIDEr and METEOR have been shown to have improved performance over other common metrics, such as BLEU.[6] In future work, further investigation of metrics will be conducted.

3. Results

The image-captioning models were scored with the automated scoring metrics CIDEr and METEOR. These metrics compared the generated captions with the reference captions for each image to produce a score reflecting the caption quality. On average, increasing the number of images in the training data set improved the performance of the models. However, as more images were added to the training data set (e.g., an increase from 1,500 images to 2,000 images), the models' performance improved gradually or remained the same.

The CIDEr scores demonstrated significant improvement in the models' performance on images of a particular object when 500 images of that object were introduced to the training data set (Fig. 1). Further additions of 500 images did not improve performance significantly. After increasing the number of images to 1,000 and later 1,500, the models improved by approximately 0.02 points on average each time. Increasing the number of training images from 1,500 to 2,000 only showed an increase of 0.001 points on average. The "cars" category score dipped at 1,000 images then increased; however, this did not demonstrate significant improvement.

The METEOR scores reflected a similar trend. Improvement between a model without training on images with a specific object and one trained on 500 images of that object was on average 0.042 points (Fig. 2). Further increases of the number of images in the training data set produced lower levels of improvement. Again, the performance of the models on images of cars showed very little improvement—only 0.022 in total. The scores of the models' performance using both the CIDEr and METEOR metrics demonstrate a similar trend of initial improvement followed by possible saturation in the scores.

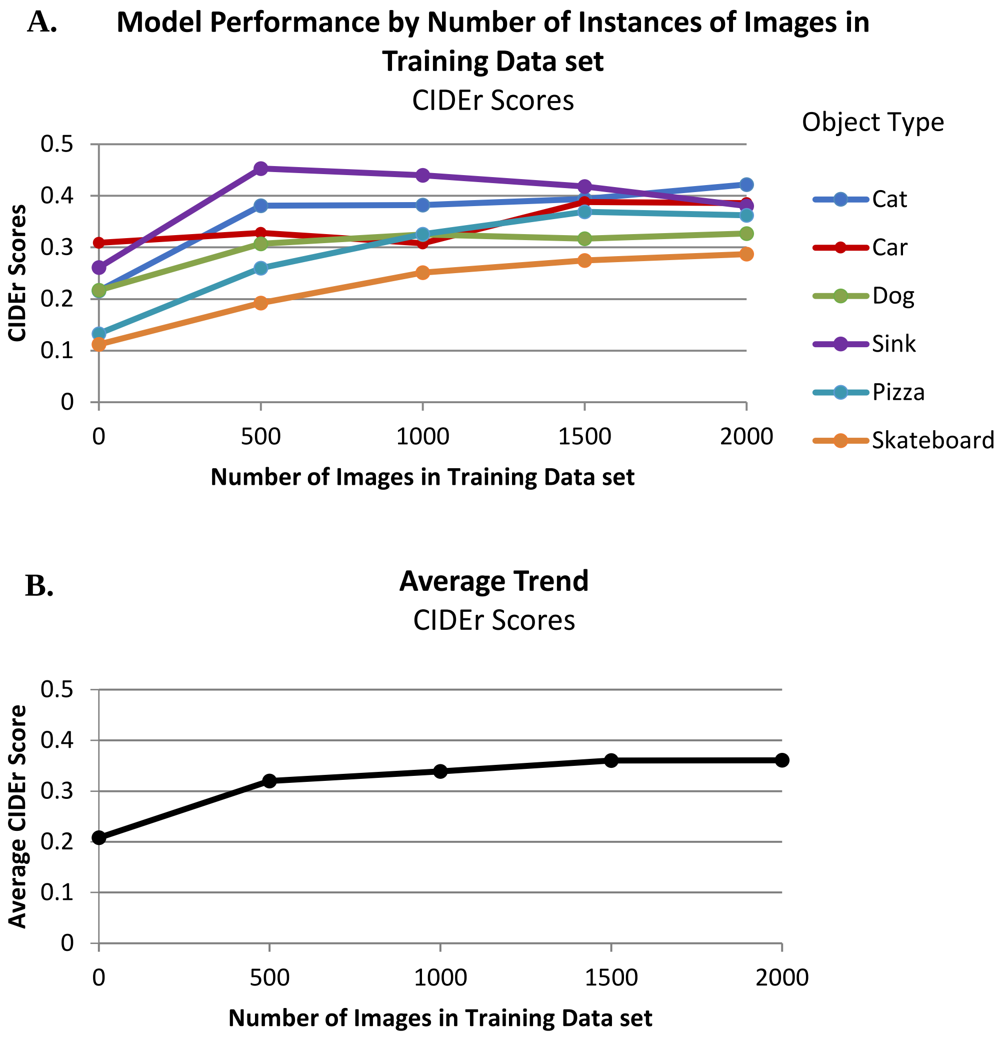

Fig. 1 Performance of the image-captioning models as measured by the CIDEr scoring metric. a) Performance of the model on images with an object (cat, pizza, etc.) vs. the number of images of that object that were included in the training data set and b) the average performance across the categories. Significant improvement was seen when the first 500 images were introduced to the training data set; however, no improvement was seen between a model trained with 1,500 images of an object and one on 2,000 images.

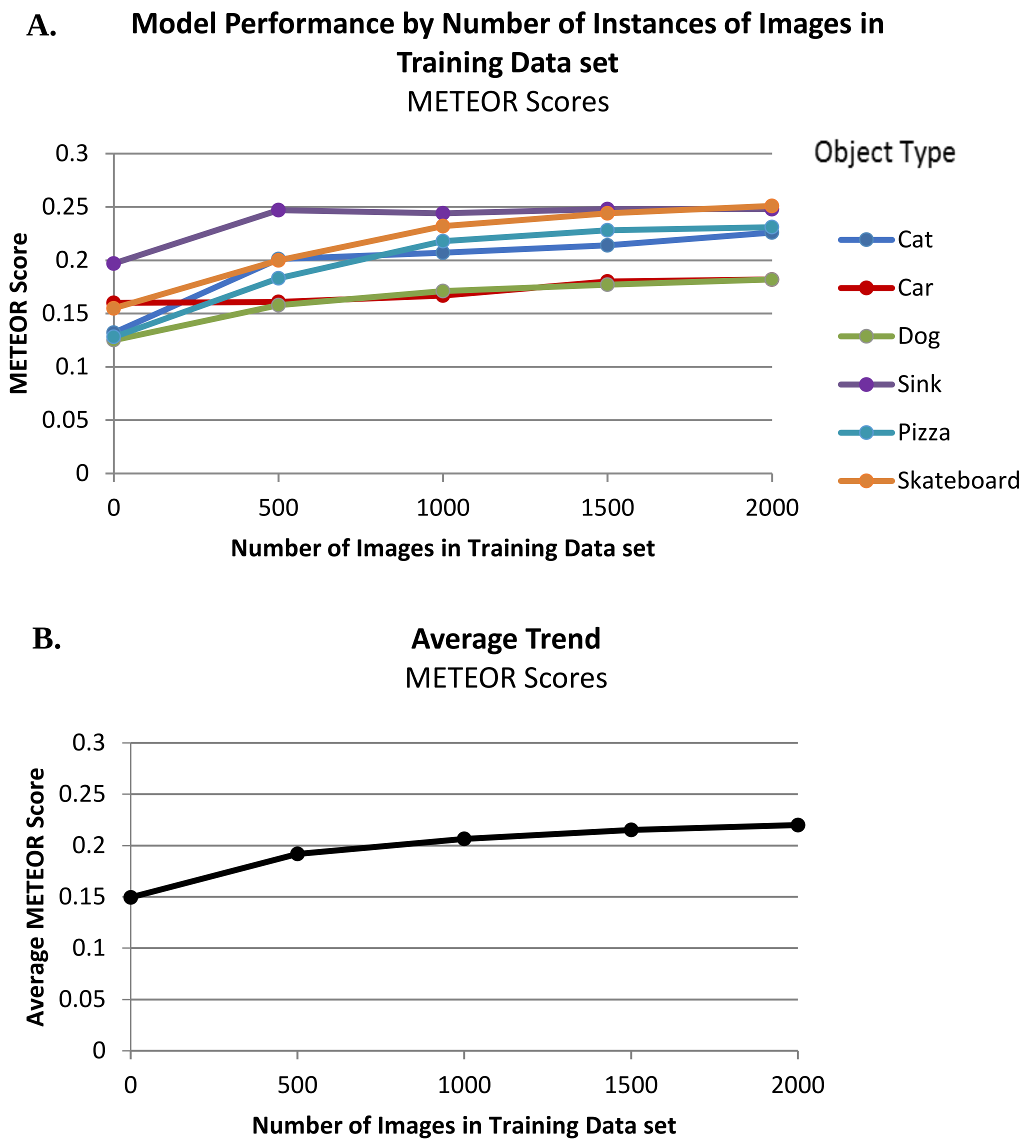

Fig. 2 Performance of the image-captioning models as measured by the METEOR scoring metric. a) Performance of the model on images with an object (cat, pizza, etc.) vs. the number of images of that object that were included in the training data set and b) the average performance across the categories. Significant improvement was seen when the first 500 images were introduced into the training data set and showed continued but less significant improvement from that point onward.

4. Discussion and Conclusion

In this experiment, the hypothesis was that increasing the quantity of images in a training data set of a particular object would improve the quality of captions generated for images of that same object. An additional goal of the study was to determine if increased training data set size improves image-captioning model performance, potentially leading to more efficient means of training high-quality models. The CIDEr scores for the models trained demonstrate improvement in model performance up to 1,500 training images. When the number of training images was increased to 2,000 images, the models did not improve, on average, and for some, performance suffered. METEOR scores corroborated this trend. The METEOR scores showed the largest improvement when the first 500 images were added to the training data set, but showed little improvement for each additional increase over 500 images. There was less or no improvement with more than 1,500 images. These initial results indicate that caption quality will saturate as more images are added to the training data set, although this saturation point may vary among objects.

The trends of model performance on specific categories are somewhat varied. For example, the "skateboard" and "pizza" categories demonstrated improvement at each step, but the growth slowed as more images were added. "Cats" showed limited improvement after 500 images and the performance of the model on images of "sinks" decreased after 500 images. "Dogs" decreased after 1,000 images. The most irregular category was "cars". Captions of images of cars showed an increase in quality, followed by a drop, and then another increase. This variability could be due to cars in images where they are not the focus of the scene. For these objects, there could be additional challenges for the model to learn meaningful captions. Overall, with the exception of cars, every other category had a point at which adding images did not significantly improve the caption quality.

While the model performance was in line with the hypothesis, more data would be necessary to draw certain conclusions about caption-quality saturation. A follow-on experiment could expand on these findings by running multiple models for each level of training images. Additionally, using this data extrapolation cannot be made on model performance with over 2,000 instances of training images of a certain type. Training additional models with over 2,000 images per category would be informative.

Possibly the greatest limitation to this experiment was caused by overfitting the models. Overfitting occurs when a model molds itself too closely to the training and validation images. It can create accurate sentences for those images, but performs very poorly on new images. Closer examination of the training data sets highlighted an imbalance of images of different objects that skewed the results to reflect the training data set. For example, the captions generated for the testing images in the "skateboard" category by the model trained on 2,000 images of skateboards were very repetitive. Out of 500 images, 225 were captioned as "a man riding a skateboard down the street". This severe overfitting of the data set could be the cause of some of the models decreasing performance as more images are added. Further experimentation would be required to determine if it was solely the overfitting that caused model improvement to plateau, or if even balanced training data sets would reach a point of maximum performance.

Future studies could expand on this research by generating a similar experiment designed to cut down on issues of overfitting. It would also be worthwhile to explore how the size and composition of training data sets contributes to overfitting. Finally, a closer investigation of how to achieve high-quality captions with low numbers of training images would be useful in situations where a large data set is unavailable. This research can be applied to optimize the training of image captioning models. It will increase the efficiency of training and provide insight into how one can train models on a smaller scale.

In addition, the results do not indicate a linear correlation to improved model performance. Rather, the growth of model performance slows down after more and more images are added to the training data set. This has potential implications for how training data sets are constructed to optimize quality, efficiency, and enhanced models for applications as diverse as image searches and human–robot interaction.

5. References

1. Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: a neural image caption generator. Ithaca (NY): Computing Research Repository; 2015 Apr 20. arXiv:1411.4555.

2. Karpathy A, Li FF. Deep visual-semantic alignments for generating image descriptions. The IEEE Conference on Computer Vison and Pattern Recognition (CVPR); 2015. p. 3128–3137.

3. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft COCO: common objects in context. Ithaca (NY): Computing Research Repository; 2014. arXiv:1405.0312v3.

4. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. Meeting of the Association for Computational Linguistics, ACL; 2001.

5. Lavie A, Denkowski MJ. The METEOR metric for automatic evaluation of machine translation. Mach Trans. 2009;23(2):105–115.

6. Vedantam R, Zitnick CL, Parikh D. CIDEr: consensus-based image description evaluation. Ithaca (NY): Computing Research Repository; 2014. arXiv:1411.5726v1.

7. Karpathy A. NeuralTalk. GitHub; 2015 Nov [accessed 2017]. https://github.com/karpathy/neuraltalk.

8. Karpathy A. NeuralTalk2. GitHub; 2015 Dec [accessed 2017]. https://github.com/karpathy/neuraltalk2.

9. Chen X, Fang H, Lin T-Y, Vendatam R, Gupta S, Dollar P, Zitnick L. Microsoft COCO Captions: data collection and evaluation server. Ithaca (NY): Computing Research Repository; 2015. arXiv:1504.00325v2.

List of Symbols, Abbreviations, and Acronyms

API	application programing interface
BLEU	Bilingual Evaluation Understudy
CIDEr	Consensus-Based Image Description Evaluation
CNN	Convolutional Neural Network
CPU	central processing unit
CUDA	Compute Unified Device Architecture
GPU	graphical processing unit
JSON	JavaScript Object Notation
METEOR	Metric for Evaluation of Translation with Explicit Ordering
MS COCO	Microsoft Common Objects in Context
RNN	Recurrent Neural Network

1 (PDF)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA
2 (PDF)	DIR ARL RDRL CIO L IMAL HRA MAIL & RECORDS MGMT
1 (PDF)	GOVT PRINTG OFC A MALHOTRA
1 (PDF)	ARL RDRL CII B A RAGLIN